Master's dissertation report

END-TO-END SLICING IN CLOUD RADIO ACCESS NETWORK (C-RAN)

Submitted by

Ravipudi Jaya Lakshmi

(204751)

Under the guidance of

Dr. Fabrizio Granelli

Associate Professor

UNIVERSITÀ DI TRENTO

Department of Information Engineering and Computer Science

University of Trento

Trento- 38123, Italy

UNIVERSITÀ DI TRENTO

Department of Information Engineering and Computer Science

Master's Degree in

Information and Communication Engineering

END-TO-END SLICING IN CLOUD RADIO ACCESS NETWORK (C-RAN)

Supervisor

Student

R. Jaya Lakshmi

Academic year 2019/2020

ACKNOWLEDGEMENT

Foremost, I would like to express my sincere gratitude and deep regard to my supervisor Prof. Fabrizio Granelli for his valuable guidance and efforts throughout the thesis work. His continued advices, skilful assessments and suggestions have improved the quality of the work in all aspects. I shall remain highly indebted for his support and motivation for the accomplishment of this work.

I also wish to express my sincere thanks to Mr. Sisay Tadesse Arzo, Ph.D. student of Prof. Fabrizio Granelli, for taking out his time to provide help from the beginning in all aspects of this work. His advice starting from the implementation to providing all the essential details to enable completion of the thesis on time is highly appreciated.

I also thank the University of Trento for giving me the opportunity and the required resources to complete the thesis.

I want to thank my parents for supporting me in all aspects of my life and for encouraging and motivating me throughout my studies. Last but not least, I also thank my friends for enthusiastically supporting me during the time of this work.

Ravipudi Jaya Lakshmi

ABSTRACT

This thesis aims to contribute a solution to the problem of end-to-end slicing in a 5G cloud radio access network (C-RAN) in the presence of three types of traffic, i.e., eMBB (enhanced mobile broadband), mMTC (massive machine-type communications) and URLLC (Ultra-reliable low-latency communications. The objective is to devise a scheduler to maximize the service admission while considering service prioritization in allocating the slice. This work proposes the use of a multi-attribute decision making (MADM) method to approach the network slicing problem and utilizes a computationally inexpensive method, Preference Ranking Organisation METHod for Enrichment Evaluations (PROMETHEE) based on the Analytic Hierarchy Process (AHP) to achieve the objective. The end-to-end network topology based on CalREN (California network) is considered and service arrival is assumed to be Poisson. The arrivals are scheduled using the PROMETHEE based on the AHP method and then considered for slicing. The network slicing finds the possible paths based on requirements vs. network constraints and determines the best route using PROMETHEE based on the AHP method in combination with minimum latency or minimum hop path policy. The results obtained in terms of admission probability, admission ratio and mean link utilization are in line with the objective and point out the possibility of using the proposed method to deal with the end-to-end network slicing problem in C-RAN.

Keywords: C-RAN, network slicing, scheduling, PROMETHEE, AHP

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1
INTRODUCTION

Communication technology has rapidly revolutionized the society in all the social, political, and economic dimensions. The thriving corporations across the globe have been rapidly adopting and using communication technologies such as network systems to enhance communication within the firms and outside their organizations. The pervasiveness and wide-spread nature of the internet and wireless technology were witnessed in the 20th century's second half. The wireless communication networks are some of the most active technologies as they are able to offer a wide range of connections with minimal to zero cables. The history of wireless communication networks has been marked by the introduction of technical standards related to 3G, 4G, and 5G [1].

The 3G or third generation network system was launched in 2001 in Japan. The wireless network standard provided an excellent mobile user experience with high-speed. The success of the 3G standards and network systems was dependent on its capacity to merge various principles of wireless networks such as the "Global Systems for Mobile Communication", "Time Division Multiple Access", and the "Code Division Multiple Access". The levels at which the 3G networks were compatible with the mobile devices were higher than the previous generations with solid internet connection speeds and the ability or capacity to support the General Packet Radio Service. Users of this wireless network technology could also switch voice calls while performing a wide range of other activities.

The dominance of the 3G technology faded as the fourth-generation mobile networks emerged. The central intention of the main challenges that were associated with 3G included the difficulty of getting 3G service by telecom operators and the requirement of multiple stations to ensure a vast network coverage [2]. TeliaSonera had launched 4G in 2010 to provide better solutions than 3G and 2G wireless networks. The central differences between the 4G and 3G networks were tighter security protocols, the data transfer rates, access methodologies, and transmission terminology. The 4G mobile meant broadband mobile services and that the devices were capable of transferring substantial data magnitudes correspondingly.

The 4G architecture is characterized by increased speed and capacity. The networks had allowed multiple devices to be connected to a single server. Moreover, the 4G architecture combines the Orthogonal Frequency Division Multiple Access, the existence of Multiple Inputs and Multiple Outputs (MIMO), Multi-user MIMO, and Long-Term Evolution (LTE). Therefore, the main ideology behind the emphasis on the migration from 3G to 4G was the need to use the session initiation protocol. While 4G’s architecture allowed it to have multiple advantages, the key disadvantages of such network systems were the difficulty to implement them, security threats to banking systems that were supported by such platforms, and their inability to handle multiple unrelated tasks [3].

The importance of information technology platforms that can facilitate several business operations forced computer scientists and other experts to develop mobile network generations with the fifth generation of network systems being the latest. The 5G is the most tailored network generation to the demands and needs of enterprises. Consequently, 5G was launched to enhance the mobile user experience, with its implementation expected to increase very faster. The transition from 4G to 5G has been rapid as the newly launched 5G provides a platform through which a majority of user needs can be addressed. The fifth-generation networks have various features such as the existence of centralized users, service providers, and network operators who are all dependent on each other. Moreover, the architecture of such systems is two-tiered. The 5G network systems are expected to incorporate the elements of cognitive radio networks, device to device communication, and cloud radio access networks. Each of these components is intended to enhance the experiences of the users while facilitating the operations of the mobile service providers and network operators. Figure 1.1 shows the cloud-based rdio access network architecture of 5G.

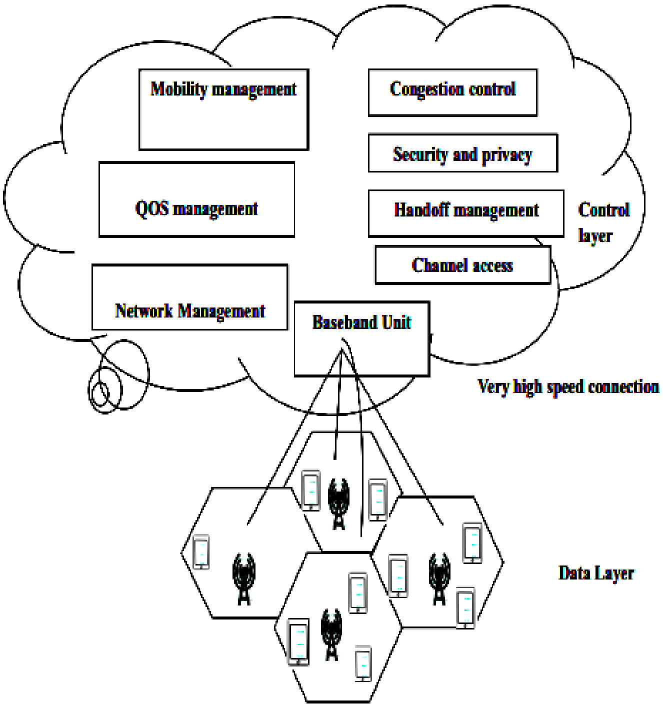


Figure 1.1: 5G Cloud-Based Radio Access Network Architecture [1]

The main components within the 5G architecture include mobility management, network management, QoS (Quality of Service) management, handoff management, and congestion control units [1]. Additionally, the 5G network has channel access, security and privacy units that fall within the control layer. A baseboard control unit manages the data layer.

The use of a cloud-based architecture and specifically the Cloud Radio Access Network (C-RAN) allows the service providers to expand their portfolios and to offer a wide range of services to their increasing number of users or clients. Moreover, the C-RAN architecture is essential as it emphasizes the concepts of service-focused resource management and scheduling and service cloud in general. As a result, the C-RAN enhances the use of computer techniques and new communication approaches that are new but tailored to the needs of the clients or consumers. The C-RAN is also a new paradigm for service providers that incorporates Wideband Code Division Multiple Access (WCDMA), LTE, and WiMAX within a Base Station structure. Therefore, the algorithms that are developed for 5G networks are less complex and highly relevant to the applications.

The 5G has been increasingly beneficial in addressing the specific challenges with cybersecurity that were prevalent in 4G networks through the provision of IP based security solutions for firms. However, 5G networks have been unable to address all needs of consumers who expect reliable and dedicated services from their network providers. Therefore, there is the need for 5G networks to undergo a redesign to enable them to offer a combination of capabilities. The central or most logical approach, from a functional viewpoint, would be designing and building dedicate networks that are adapted to solve specific consumer needs [4]. Such networks would be beneficial in that they would have a tailor-made functionality that permits them to address a single need at a time. However, from an economic viewpoint, such networks would be costly and turn away potential customers and investors.

The is insufficiency of information on scheduling algorithms and admission controls and specifically in the context of 5G networks. The projection and anticipation of 5G network providers were initially to support a wide variety of services that match different requirements [5]. Specifically, the mobile broadband usage circumstance and guaranteed bit rates were to be increased substantially to match the dynamic service requirements. The central goal of network slicing is to provide a mechanism through which 5G algorithm scheduling can be initiated, and admission control be enhanced to increase the security requirements of such network systems.

The network slicing is defined as an approach or concept that entails the operation of several logical networks that are viewed as business operations that are virtually independent on a single physical infrastructure [6]. The 5G networks have long been perceived as the communication technologies that have the potential for permitting business customers to enjoy data processing and connectivity that are customized to their needs and specific business requirements. Furthermore, a network slice is crowned as an independent end-to-end logical network, meaning that such a network runs on shared infrastructure and has the capability of addressing specific sets of needs requested by consumers. Network slicing is highly applicable to the 5G mobile systems as it targets all the three types of communications ("ultra-reliable low-latency communications (URLLC)," "massive machine-type communications (mMTC)," and "enhanced mobile broadband (eMBB)".

The need for optimal network slicing is more pronounced in 5G networks as compared to other generations of networks [7]. The intervals of time between the response and stimulation processes in a 5G C-RAN are generally high after network slicing. Moreover, network slicing increases data security and energy efficiency of 5G networks. The drive to introduce the concept of network slicing is convenience in accessing data within a cloud system. The network slicing enables reachability through the quick access of data in such systems, massive connectivity as multiple networks share the same virtual infrastructure, mobility and throughput.

The incoming 5G services should be scheduled in an optimal manner and the network resources should be allocated optimally depending on the requirements of the service coming. There is a need to enhance the knowledge regarding optimal end-to-end slicing in 5G C-RANs.

The thesis is arranged in six chapters. Chapter 1 gives the introduction. Chapter 2 presents the literature review on scheduling and slicing strategies and the use of heuristics and multi-attribute decision making (MADM) methods to solve the network slicing problems. Chapter 3 describes the research problems considered and the methods used. Chapter 4 presents the details of implementation of the scheduling and the network slicing processes. Chapter 5 presents the results and discussion. Chapter 6 concludes the thesis, summarizes the main findings and outlines the scope for future work.

Chapter 2

LITERATURE REVIEW

Proper control strategies are required when there is an imbalance between demand and supply in any kind of situation. Analogously, an efficient admission control algorithm is essential for the proper management of the available network resources. The physical network is owned and managed by the infrastructure providers (InPs) and the virtual network operators (VNO) sell their services by using the network resources of the InP. The InP has the decision making power to admit or reject the slice requests based on what is being asked by the slice and what network resources are available in the resource pool. The service chain embedding or virtual network embedding (VNE) succeeds slice admission. The utilization of the network is dependent on the efficiency of the admission control module, which is the decision process of how and when to allow the slice requests to get access to the network resources. Since the present work focusses on the efficient scheduling and slice provisioning and not on the VNE problem, the literature review in this chapter will only include research works about the scheduling and slicing.

Foukas et al. [4] conducted a survey on network slicing in 5G and discussed the challenges. Guo and Suárez [5] proposed a framework for slicing of a 5G RAN. The earliest deadline first (EDF) scheduling was used for radio resource allocation in RAN slicing. The main requirements of RAN slicing were fulfilled by the scheme. Jia et al. [6] proposed a caching resource sharing scheme to efficiently share the limited physical caching resource of InP among multiple network slices. The authors had formulated the caching resource sharing problem as a non-cooperative game, and the solution was obtained by an iteration algorithm. Tang et al. [7] proposed a slice-based virtual resources scheduling scheme with non-orthogonal multiple access technology to increase the quality of service of the system. The power granularity and subcarrier allocation strategies were formulated into a Markov decision problem, aiming to maximize the total user rate.

Any slice admission algorithm is designed based on some strategy. The strategy may be random, priority-based, first-come-first-served (FCFS), greedy, or semi-greedy. Han et al. [8] applied the FCFS strategy to a multi-queueing system accommodating heterogeneous requests to study the admission control problem by deriving a statistical behavior model and proposed a utility-based admission control optimization. The performance metrics used by the authors were the admission rate, overall network utility and the average waiting time of the request. The algorithm can be used with any of the strategies listed before. Still, the results would be dependent on the selection of strategy, thereby initiating the necessity of using strategy optimization. FCFS strategy is simple and straightforward but rejections will happen if any request arrives at a time simply because no resources are available. There is no optimality and no novelty since the FCFS strategy follows the natural order in which the requests/slices arrive.

Slice admission can work upon priority strategy i.e. the slice provider can decide to admit slice requests preferentially like Ultra-Reliable Low-Latency Communications (URLLC) requests, which have strict latency requirements. Raza et al. [9] proposed Reinforcement Learning (RL) for profit optimization keeping in mind the latency requirements in Radio Access Network (RAN). A neural network received the input requests and decided whether to accept it or not based on the potential reward it could get. The authors had reported that the approach increased the performance by more than 50% but at the expense of selective rejections of low revenue services. Therefore, this approach might be beneficial from the InP's point of view but it is biased towards the VNOs that buy the high priority slices, which is not right in terms of fairness. Moreover, it might not be sufficiently profitable to InPs as well since the occurrence of such high priority requests, e.g. autonomous vehicles, is scarce when compared to the low-revenue generating requests which occur more regularly, e.g. broadband requests.

To avoid InPs' preference for selective services, a partially greedy algorithm can be considered as the one proposed by Challa et al. [10]. The authors had implemented a partially adaptive greedy algorithm (PAGE) considering both profit maximization as well as valuing the service level agreements (SLA) of all kinds of requests. The partially greedy algorithm allows the low profit requests also to be explored with a small probability. An analogy of slice admission was made to the knapsack problem and PAGE was applied to the combinatorial integer programming model to obtain a sub-optimal solution. The problems which use greedy or partially greedy algorithms normally get the sub-optimal solutions.

Optimization techniques are used to improve the performance of the slice admission control algorithms to have more solution exploration and exploitation capability. The objective function to optimize differs from situation to situation as per the requirement of the InP, such as maximizing revenue, minimizing congestion, etc. Bega et al. [11] designed an algorithm to learn the admission policy autonomously while ensuring that the slice requestors are satisfied. The authors had first described a value iteration method to capture the best policy but owing to its high computational load, a neural network based algorithm was proposed. The algorithm was tested for both small and large scale scenarios with elastic and inelastic traffic and its superiority was proved over the Q-learning model and random strategy model. Generally, the optimal algorithms are mostly complex when compared to the FCFS, priority or greedy ones since the search process is more extensive and hence proper design and tuning are required to counteract the computational complexity.

Many papers had been published in the literature for network slicing using heuristic methods. The domain of slicing can be radio access network (RAN), core network, cloud network or end-to-end. Different researchers had proposed and applied different heuristic based and machine learning based slice admission algorithms in each of the different slicing domains. Anand et al.[12] derived an optimal scheduler for joint scheduling of URLLC traffic and enhanced Mobile Broadband (eMBB) type traffic. The physical layer was considered and the aim was to find the optimal rate allocations to include prioritization concept for URLLC type and maximization of utility for eMBB type. The resources were given to an eMBB request at the slot boundary of the resource block and at mini-slot (part of a slot) boundary for URLLC type request even if some eMBB request was ongoing in that particular slot. The users would perceive the change of rates of the requested service due to this overlapping of slots. Therefore, the authors had formulated this problem as a joint scheduling optimization type. Convex and threshold loss models were used to study the impact on the eMBB rates due to URLLC. Interrupting eMBB services just for URLLC type every time is not fair and therefore, a sharing factor (β) was introduced such that in a given mini-slot, only $(1- \beta)/ \beta$ units of URLLC were served and the remaining were queued until the opportunity of getting next mini-slot.

Korrai et al. [13] addressed the mutual coexistence and simultaneous resource allocation problem of URLLC and eMBB traffic in the context of time-frequency resource blocks (RB) in a downlink Orthogonal Frequency Division Multiple Access (OFDMA) situation. Like in [12], here also the maximization of rate was focussed upon but also including the slice isolation constraint. But the approach employed was different and an adaptive modulation coding was used in this work. The optimization problem was formulated as a combinatorial type and since it was a NP-hard problem, it was converted to a continuous linear program and was solved using the standard CVX tool. It is to be noted that unlike the mini-slot approach in [12], here, an ongoing eMBB transmission was not interrupted. Constraints were put for the optimization such that a particular RB was assigned to a single user of any one traffic type. But at the same time, to guarantee the latency constraint of URLLC type, a condition was put such that it received at least one RB for every N-slots.

It is not always necessary that the objective function has to be maximization of utility or data rate but can vary depending on the designer requirement. For example, Ginige et al. [14] maximized the number of users i.e. maximized admissions of eMBB type with a guaranteed bit rate while also prioritising all URLLC users in a single cell multiple input single output (MISO) system. Unlike the puncturing approach in [12] or the guaranteed RB per N slots approach of [13], here the possible supportable number of eMBB users was found out first while satisfying the reliability and latency requirements of URLLC users. The objective function to increase the cardinality of the accepted eMBB users was formulated as an l_0 NP-hard problem which was solved using l_0 approximation and sequential convex programming. The optimization parameters considered were bandwidth, power and beamforming directions and the results showed that the algorithm achieved the aim but the solution was a sub-optimal one considering the nature of the problem. The algorithm can be extended to multi-cell scenarios.

Albonda and Perez-Romero [15] studied the radio access network slicing and a slicing scheme based on reinforcement learning and a heuristic algorithm was proposed. The input traffic types considered were eMBB and traffic from vehicle-to-everything (V2X) with the aim of maximizing resource utilization while maintaining congestion-free network so that the varied requirements of the traffic were fulfilled. Unlike most of the other works available in the literature and the ones mentioned in this chapter, the authors had addressed all three directions of traffic movement i.e., uplink, downlink and sidelink. The results were validated against two reference schemes and with the reinforcement learning approach. It was found that the proposed solution managed to achieve better resource block (RB) utilization with improvements in bit rate. However, the authors had not considered the multiple traffic types in a single slice, which is a better realistic scenario and can be used as a possible application of the algorithm presented.

Dietrich et al. [16] described the aspects of network function placement on virtualized Long-Term Evolution (LTE) core networks. Sattar and Matrawy [17] focussed on the 5G core (virtual Evolved core) network slicing while keeping in mind the intra-slice isolation (physical isolation between Virtual Network Functions (VNF) of a slice) and end-to-end network delay. The authors had extended the work reported by Dietrich et al. [16]. The original work in [16] was focussed on LTE core networks but Sattar and Matrawy [17] used the idea to apply to 5G core network. The objectives for the 5G core network were to guarantee the end-to-end delay, give intra-slice isolation, and to find a minimum delay path for the slices by formulating a mixed-integer linear programming (MILP) function. The algorithm was modelled using A Mathematical Programming Language (AMPL) and CPLEX 12.8.0.0 was used to solve the MILP problem. Among the possible slice solutions, the path with the least utilized servers and the lowest delay was assigned to the incoming request. The notable point was that only one requirement i.e., the end-to-end delay was considered by the authors. Hoever, including more specifications to meet varied demands from diverse applications would make the slicing solution more efficient. Also, the computational time of the proposed algorithm increases with the intra-slice isolation (reliability) demand.

Most of the slice algorithms focus on either RAN part or Core part and the research works considering end-to-end slicing are comparatively less. Also, most of the studies only consider the combination of eMBB and URLLC for scheduling decisions whereas the third use of 5G, massive Machine Type Communications (mMTC), is equally important since it encompasses all the communications of Internet of Things (IoT).

Nakao et al. [18] summarized their research efforts on 5G mobile networking to enable end-to-end network slicing and discussed the application use cases that should influence the designs of the network slicing infrastructure. Vassilaras et al. [19] focused on the algorithmic challenges that emerge in efficient network slicing, necessitating novel techniques from the disciplines of operations research, networking, and computer science. The works of Nakao et al. [18] and Vassilaras et al. [19] are the two important works detailing the theoretical and algebraic aspects of end-to-end slicing.

Kammoun et al.[20] considered end-to-end slicing for all three 5G use cases based on availability, reliability and delay parameters. A simple low cost heuristic algorithm to solve an exponential utility function was proposed. A path score was computed using the exponential utility function and the path with the highest score along with the computing nodes was allocated for the requested service. It was claimed that the proposed algorithm was better than the ones that only consider delay or availability or reliability since this one considered every criterion by giving equal weightage to all. Because it is not necessarily true that the least latency path would be the reliable one too. Therefore, it becomes important to consider all the criteria while allocating a slice. It is noted that the algorithm's simplicity comes with the drawback that it does not really take into consideration the dynamicity of the network resource pool and the priority or latency sensitivity of the services is also not considered. Bassoli and Granelli [21] proposed a novel general model for network slicing based on multilayer graphs, linear algebra and algebraic graph theory. The proposed framework had generalised specific legacy models.

The research works discussed until now are based on heuristic and machine learning methods and the researchers had optimized the slicing specific parameters to achieve a fixed goal(s). However, the solutions were sub-optimal and at a time, any one type of service would be at some disadvantage due to the importance given to a specific-criterion, which is not favorable to it. The literature shows that the researchers had used some techniques falling under multi-attribute decision making (MADM) field to solve the network slicing problem. The MADM methods are used to choose a right alternative from a finite number of alternatives based on the data or features of each attribute with respect to every alternative.

Bakmaj et al. [22] used an MADM method known as TOPSIS (Technique for Order Preference by Similarity to Ideal solution) for an optimal network instances selection. The authors had considered the criteria of throughput, Quality of Service (QoS) level, security level and cost of service. The aim was to select the network slice by ranking the possible slices by simultaneously considering the four criteria. The authors had modified the original TOPSIS method to solve the rank reversal problem and also proposed alternative ranking techniques. Rosa and Rothenberg [23] also used TOPSIS method for choosing the better slicing alternatives.

Li et al. [24] combined a heuristic-based method and an MADM method known as VIKOR (VIšekriterijumsko KOmpromisno Rangiranje) for a two-stage slice provisioning in core networks. The network nodes were evaluated and ranked by the VIKOR method. These nodes were subsequently sliced. A strategy to improve the slice admissions was proposed. A path factor was defined to make sure that the slice is allotted with the path that has the lowest bandwidth utilization as well that offers a minimum number of hops to the destination. The algorithm was tested under varying load scenarios with different security requirements.

The TOPSIS method used by Bakmaj et al. [22], Rosa and Rothenberg [23] and the VIKOR method used by Li et al. [24] are computationally intensive and difficult to understand. Furthermore, there is no provision in these methods for systermatically deciding the weights of the criteria or attributes. The usage of MADM methods in 5G service scheduling have been far less when compared to their heuristic counterparts. The MADM methods had been used for finding the best network for heterogeneous wireless networks in 4G domain [25 - 29], but very few applications are there with respect to 5G domain. Hence, there is a need to apply the simple and powerful MADM methods in 5G domain.

It is noticed from the literature review that most of the research works concentrated on the joint scheduling of only eMBB and URLLC type of traffic. However, scheduling the two services in the presence of mMTC type of communications are not studied or rarely studied in the literature. For the most part, the resource allocation algorithms considered the physical layer in terms of slot allocation in the resource blocks (RBs) in the downlink direction. Whereas research works about network layer slicing are far less when compared to the physical layer ones. With regard to the slicing domain, RAN slicing is mostly considered in comparison to the core network slicing or end-to-end slicing. Also, the predominantly used approaches are heuristic-based and the solutions using MADM methods are not widely explored in 5G slicing scenarios. Thus, there is a need to research more on the areas mentioned above because all three types of traffic can arrive simultaneously in a real network. The three types of traffic demand and compete for resources belonging to all parts of the network to reach the destination efficiently rather than in a specific RAN or core or data center component.

Therefore, the objectives of the current work are framed as below.

1. To devise an optimal scheduler for optimal slice allocation with the help of an improved MADM method known as improved PROMETHEE (Preference Ranking Organization Method for Enrichment Evaluations) for better examination of the conflicting demands of the diverse service requests.
2. To increase the number of admitted services and the mean link utilization.

The next chapter presents the research problems considered and a detailed description of the improved PROMETHEE method.

Chapter 3

RESEARCH PROBLEMS CONSIDERED AND THE METHODS USED

This chapter presents the research problems considered in the present work and the methods used to solve the same.

3.1 Research Problems Considered

Scheduling is an essential component of any system that needs to satisfy the demands of incoming processes in terms of sharing common resources. As discussed in the literature review in Chapter 2, the scheduling schemes proposed by various researchers took into account some specific strategies of scheduling, such as FCFS, priority-based, greedy, semi-greedy, etc. All these approaches have proven to have their own merits and demerits. When the FCFS technique is considered, services with higher bandwidth occupies more significant parts of the capacity thereby starving the ones with lower bandwidth. The same can be said about the priority-based approach since the top priority services are allowed most of the time. Therefore, there is no balance of services in the network to provide a fair and acceptable admission probability and or admission rate. It is henceforth required to find a satisfactory solution to all the problems encountered in the other methods.

In addition to addressing the scheduling schemes, another problem to address is the end-to-end slicing. Although an excellent and sophisticated technique of scheduling will provide better functionality, there is also the need for an optimal slicing. End-to-end slicing takes into consideration all the links and nodes at the various sectors of the network. A service can be dropped at any stage of the network if the slicing approach is not optimal. In other words, if the conditions of the slice at the RAN part are met, it does not guarantee that the same conditions could be satisfied at the core part of the network. A node or link failure can also result in services being dropped if the slice is not chosen well.

These problems are attempted in the present work using MADM methods. The MADM methods are applied to schedule as well as to allocate the slice. For selecting the best slicing alternative, an MADM method is used in conjunction with the concept of minimum delay path and minimum hop path. The next section briefly describes the MADM methods.

3.2 Multi-Attribute Decision Making (MADM) Methods

Taking decisions is an integral aspect of any person's life. It can be as trivial as choosing an attire to as crucial as selecting a course of action for a project/industrial task, which would yield maximum benefit/productivity. It is not limited to humans, but even machines encounter situations when it is needed to decide the best possible action or select the best viable option given some rules and criteria. For example, autonomous cars need to make quick decisions based on the real-time conditions of the traffic around them. Making decisions out of intuition for simple problems may be considered acceptable. But when risks are high and since the chosen option can affect the outcome both positively or negatively, it is essential to evaluate multiple criteria and use scientific methods to structure the problem correctly and arrive at a well-informed decision.

Multi-criteria decision making (MCDM) is a popular branch of decision making that falls under the discipline of Operations Research. It is categorized into multi-attribute decision making (MADM) and multi-objective decision making (MODM), depending on the domain of the alternatives. MODM methods are preferred when the problem is a decision problem, and the decision variable has values in the continuous or integer domain. These methods generally have an infinite or a large number of alternatives. However, the MADM methods are applied when faced with the problem of selecting the best viable option (or sorting and ranking) from a given finite set of alternatives, which are influenced by several criteria or attributes [30, 31]. All the elements in the decision table must be normalized to the same units, so that all possible attributes can be considered.

There are various MADM methods available in the literature. To name a few, simple additive weighting (SAW) [32], Analytic Hierarchy Process (AHP) [33], Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)[34], Compromise Ranking Method (VIKOR) [35], Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE) [36], etc. The PROMETHEE method combined with AHP is used in the present work and is explained in detail in the following section.

3.3 Preference Ranking Organization METHod for Enrichment Evaluations (PROMETHEE)

The PROMETHEE method belongs to a category of outranking methods and was introduced by Brans et al.[36]. The steps of PROMETHEE are explained as follows.

Step 1: Identify the 'proper' attributes (or criteria) for the problem. A quantitative or qualitative value or a range of values may be used to decide the inclusion of a particular attribute as the 'proper' one. Short-list the alternatives based on the attributes identified. The attribute values can be the already available data or the estimated data by the decision-maker. Table 3.1 shows a sample decision matrix formed based on these alternatives-attributes. Let a_i (for i=1,2….X) denotes i^{th} alternative, b_j (for j=1,2….Y) denotes j^{th} attribute, p_{ij} (for i=1,2….X; j=1,2….Y) are the performance measures i.e. values of alternatives with respect to each attribute, and w_j (for j=1,2….Y) denotes the weightage assigned by the decision maker to the j^{th} attribute. Table 3.1 shows the general form of the decision table used in MADM methods.

Table 3.1 General Form of Decision Table in MADM [30]

Alternatives			**Attributes**			
	b_1 (w_1)	b_2 (w_2)	b_3 (w_3)	- (-)	- (-)	b_Y (w_Y)
a_1	p_{11}	p_{12}	p_{13}	-	-	p_{1Y}
a_2	p_{21}	p_{22}	p_{23}	-	-	p_{2Y}
a_3	p_{31}	p_{32}	p_{33}	-	-	p_{3Y}
-	-	-	-	-	-	-
-	-	-	-	-	-	-
a_X	p_{X1}	p_{X2}	p_{X3}	-	-	p_{XY}

Step 2: Decide the attributes' weights

Brans et al. [36] did not mention any specific method to assign the weights to the attributes. Therefore, to find out the weights of the attributes, the present work uses the AHP method and the same weights will be used in the PROMETHEE method. The procedure for finding the weights is outlined below.

Step 2.1: Find the relative importance of attributes

A matrix consisting of attributes on rows and columns is formed and the matrix values are filled using the relative importance of each attribute over another attribute. The diagonal values are all 1's since an attribute compared to itself in these matrix positions. The other positons of the matrix are filled with values showing importance of a particular attribute over another using a standard scale rather than filling with random intuition. Table 3.2 shows the relative importance scale.

Table 3.2 Relative Importance Relations Table in AHP [33]

Degree of importance	**Definition**
1	Two attributes have equal importance
3	One attribute has slightly more importance than the other
5	One attribute has strongly more importance than the other Strongly more important
7	One attribute has very strongly more importance than the other
9	One attribute has absolute more importance than the other
2, 4, 6, and 8	Intermediate values

Matrix M1 shows a sample pairwise comparison matrix with Y number of attributes. r_{pj} ($p=1,2\ldots Y$; $j=1,2\ldots Y$) is the importance value of p^{th} attribute relative to the j^{th} attribute and $r_{jp}= 1/r_{jp}$ and $r_{pj}=1$ if $p=j$.

$$
M1_{YxY} = \begin{array}{c|cccccc|}
\text{Attribute} & 1 & 2 & 3 & --- & --- & Y \\
\hline
1 & r_{11} & r_{12} & r_{13} & --- & - & r_{1Y} \\
2 & r_{21} & r_{22} & r_{23} & --- & --- & r_{2Y} \\
3 & r_{31} & r_{32} & r_{33} & --- & --- & r_{3M} \\
--- & --- & --- & --- & --- & --- & --- \\
--- & --- & --- & --- & --- & --- & --- \\
Y & r_{Y1} & r_{Y2} & r_{Y3} & --- & --- & r_{YY}
\end{array} \qquad (3.1)
$$

Step 2.1: Relative normalized weights (w_j)

The normalized weights for each attribute is found out by (a) computing the geometric mean of each p^{th} row of matrix M1 using Eq. (3.2) and (b) normalizing the obtained geometric means using Eq. (3.3).

$$GM_j = \left(\prod_{j=1}^{Y} r_{pj}\right)^{1/Y} \tag{3.2}$$

$$w_j = \frac{GM_j}{\sum_{j=1}^{Y} GM_j} \tag{3.3}$$

The geometric mean is simple to use and makes it easier to find the Eigen value (in the next steps).

Step 2.2: Compute matrices M3 = M1*M2 and M4= M3/M2 where M2= $[w_1, w_2, \ldots, w_Y]^T$. The matrix M2 is the vector of weights found in step 2.1.

Step 2.3: Find the maximum Eigen value (λ_{max}) which is the average of matrix M4.

Step 2.4: Compute the consistency index CI = $(\lambda_{max} - Y)/(Y - 1)$. The smaller the CI, the smaller will be the deviation from the consistency.

Step 2.5: Compute the consistency ratio CR = CI/RI, where RI is the random index. RI is a standard value defined as per the number of attributes. Table 3.3 shows the RI values for different number of alternatives [33].

Table 3.3 RI Values for Different Number of Attributes [33]

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RI	0	0	0.52	0.89	1.11	1.25	1.35	1.4	1.45	1.49	1.51	1.54	1.56	1.57	1.59

The CR is a significant value, which is acceptable only if it is less than or equal to 0.1, i.e. the error in judgements is less than 10%. The CR is a measure of how much consistent the relative importance relations are in matrix M1 and reflects the correctness of the judgment with regard to the problem under consideration. The decision maker can accept the weightages of the attributes obtained by this AHP method, if the value of CR is less than 0.1.

Step 3: Preference function calculation (P_j)

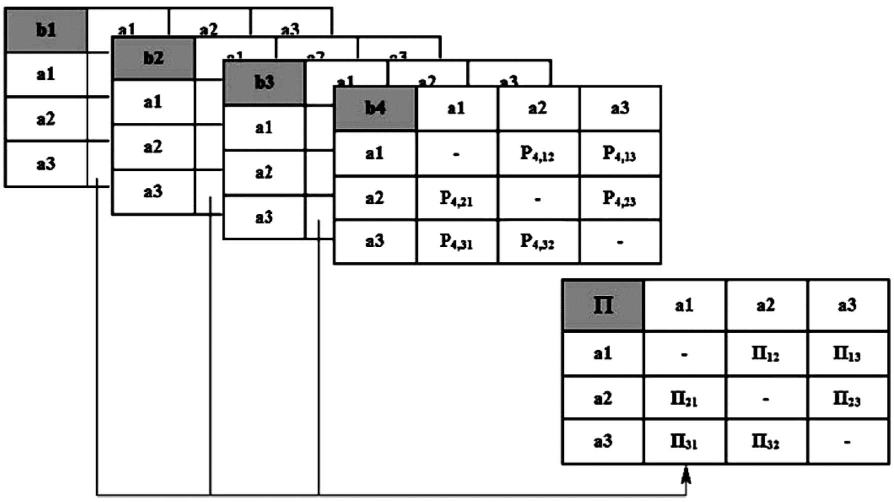
This step requires information about the preference function (P_j) desired by the decision-maker to assess the contribution of a given alternative concerning a given attribute. $P_{j,a1a2}$ is an equivalent representation of the evaluation difference between two alternatives ($a1$ and $a2$) with respect to a particular attribute. It outputs a preference degree ranging from 0 to 1.

$$P_{j,a1a2} = G_j\left(b_j(a1) - b_j(a2)\right) \tag{3.4}$$
$$0 \leq P_{j,a1a2} \leq 1$$

The PROMETHEE method offers six different types of preference functions. The "usual" type preference function is the simplest one and is used in this research work. It is nothing but the difference between the attribute value b_j for the alternatives $a1$ and $a2$. For example, for a problem with 3 alternatives and 4 attributes the matrices denoting the differences between the alternatives for each attribute will look like the one shown in Figure 3.1. Then, the multiple attribute preference index is computed using Eq. (3.5), i.e. the weighted average (w_j, j=1,2,…,Y) of the preference functions (P_j).

$$\prod\nolimits_{a1,a2} = \sum_{j=1}^{Y} w_j P_j \tag{3.5}$$

The $\prod_{a1,a2}$ indicates the preference intensity of the decision maker of one alternative $a1$ over another alternative $a2$ and its value varies between 0 to 1.



$$\Pi_{31} = \sum_{j=1}^{4} w_j P_{j,31}$$

Figure 3.1 Computation of Preference Indices for a Sample Decision Making Problem With 4 Attributes and 3 Alternatives [37]

Step 4: Calculation of leaving, entering flow and net flow

$$\Phi^{+}(a) = \sum_{x \varepsilon A} \Pi_{xa} \tag{3.6}$$

$$\Phi^{-}(a) = \sum_{x \varepsilon A} \Pi_{ax} \tag{3.7}$$

$$\Phi(a) = \Phi^{+}(a) - \Phi^{-}(a) \tag{3.8}$$

Where, $\Phi^{+}(a)$ is called the leaving flow, $\Phi^{-}(a)$ is called the entering flow, and $\Phi(a)$ is called the net flow of the alternative *a* belonging to a set of alternatives A. The leaving flow value is a measure of how much the particular alternative dominates the rest of the alternatives and the entering flow value denotes how much the alternative is being dominated by rest of the alternatives. The net flow is a value function whose higher value is desired. Higher the value of net flow of an alternative, higher is that particular alternative's attractiveness. Therefore, the net flow values are ranked by giving the best rank to the highest net flow value.

An example is included in APPENDIX 1 to demonstrate the step-by-step working of the improved PROMETHEE method combined with the AHP.

The next chapter presents the implementation details of scheduling and network slicing concepts used in this thesis.

Chapter 4

IMPLEMENTATION OF THE SCHEDULING AND THE NETWORK SLICING PROCESSES

The aim of the present work is to define an optimal scheduler for the three types of 5G use case services, namely enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC) and massive Machine-Type Communications (mMTC) so that the end-to-end network slicing is optimal by taking into account the presence of high priority sensitive services but also taking care that the flexible services are not starved while considering all their requirements and the available network resources.

The slicing domain considered is end-to-end in uplink direction i.e. each network slice is composed of links between radio access network (RAN), microwave backhaul, fiber backhaul and data center network. As an illustration, Figure 4.1 shows a good visualisation of end-to-end network slicing for the three service types. Each service might ask for different requirements and virtual network functions (VNFs) and therefore the connection topology varies. Because the slicing is restricted to the network layer i.e. layer 3 of the Open Systems Interconnection (OSI) model, the task is to find the optimal paths/links for routing the arriving slice requests.

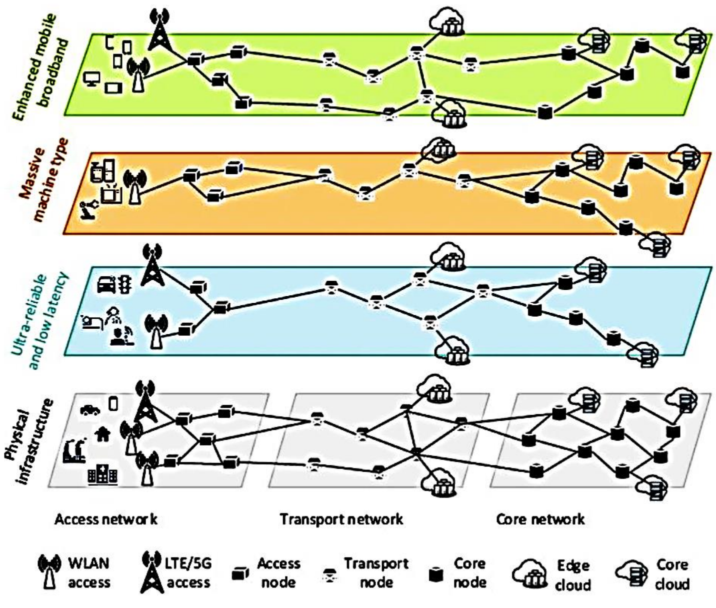
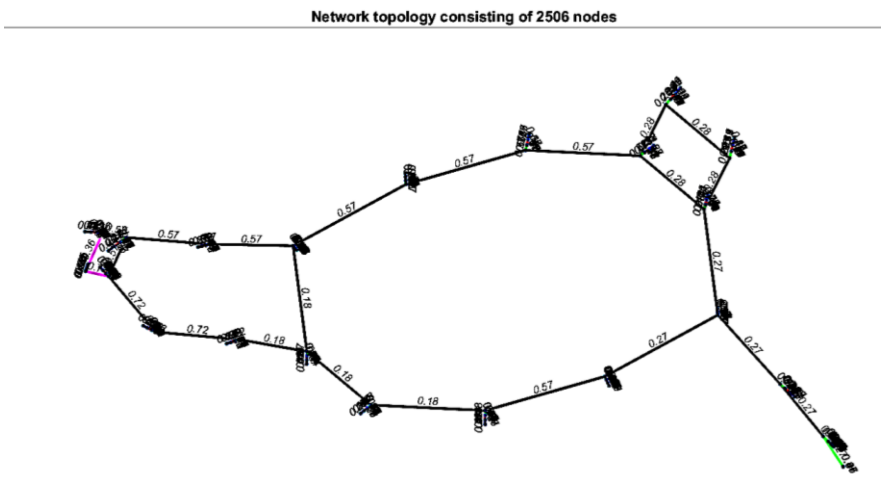
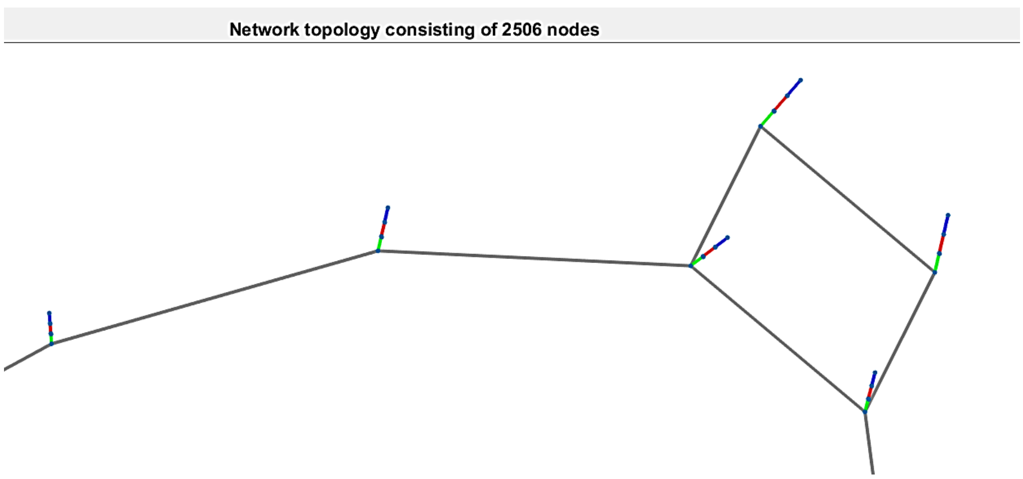


Figure 4.1 Illustration of End-to-End Network Slicing

The cloud radio access network (C-RAN) topology is emulated in MATLAB as a loopless hypergraph. The backbone network is represented using the California network (CalREN) and is extended to create the full end-to-end network topology. Figure 4.2(a) shows the generated network topology with a total of 2506 nodes. Since the number of nodes is a large number, the figure produced by MATLAB is not clear enough to distinguish the RAN, microwave, fiber and data center parts. The jumbled up numbers in Figure 4.2(a) are the weights of the graph edges which in the present case are the delay values of those particular links in milliseconds. To prove that all the nodes are actually generated as claimed, a zoomed-in version of Figure 4.2(a) is shown in Fgure 4.2(b). The different colors represent different parts of the network but still the resolution is not high enough to individually distinguish the links.



(a)



(b)

Figure 4.2(a) End-to-End Network Topology With Edge Weights; **(b)** Zoomed-in Version of Figure 4.2 (a) (Without Weights for Better Visualisation)

Let there be a graph, G =(X,Y) composed of a set of vertices/nodes (X) and a set of edges (Y) denoting the physical network of 5G as shown in Figure 4.3.

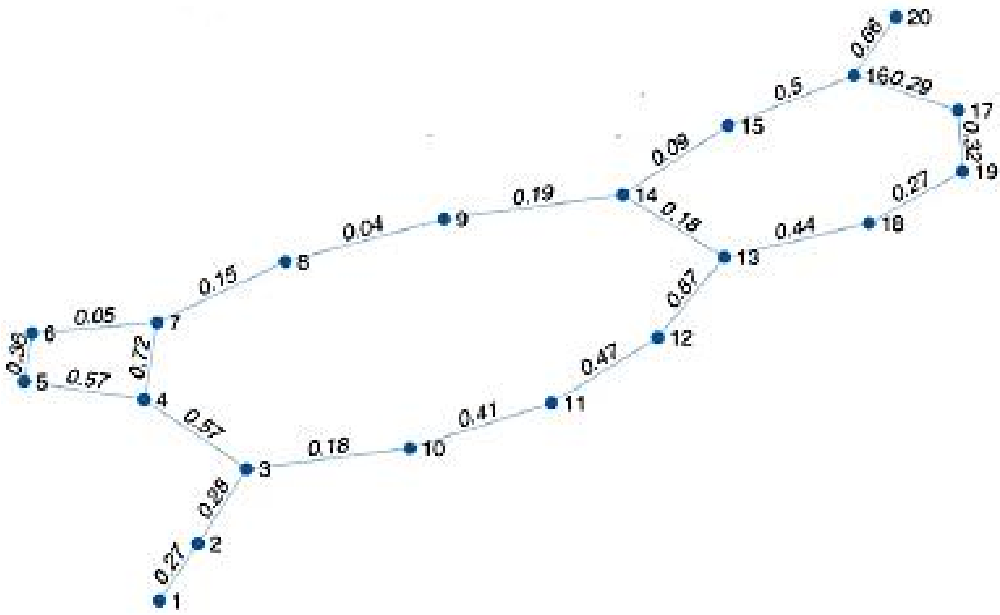


Figure 4.3 A Sample Network Undirected Graph [21]

Let $A_G = (a_{ij})$ be the adjacency matrix of size |X|*|X|. The adjacency matrix is a matrix of 1's and 0's. If $a_{ij}=1$, it means there exists a connection/link between the i^{th} node and the j^{th} node, else it is 0, and $a_{ij} = a_{ji}$. The adjacency matrix is used as a reference matrix to initialise all connections with the parameters (also called attributes or criteria) of capacity (C), latency (T) and reliability (F) and we have C_G, T_G and F_G matrices accordingly. The delay matrix is calculated using Eq. (4.1).

$$T_G = T_P + T_R \tag{4.1}$$

where T_P is the propagation delay (constant values) and T_R is transmission delay which is inverse of the available link capacity at a particular time. The total end-to-end delay of a path is calculated by summing up the delays of the links in that path. The reliability matrix is given by Eq. (4.2).

$$F_G = \rho_{ij} \tag{4.2}$$

$$\rho = 1 - P_f \tag{4.3}$$

where, P_f is the failure probability of the link and ρ is the success probability. The total end-to-end reliability of a path is computed by taking the product of all individual reliabilities of the links in that path. Table 4.1 shows the parameter values of different parts of the network.

Table 4.1 Parameter Values of Different Parts of the Network

C-RAN components	Latency (ms)	Reliability (%)	Link capacity (Gbps)
RAN	1	≈100	0.5/1
Microwave backhaul	0.65	99.3	1/10
Fiber backhaul	1	≈100	500/1000
Data center network	0.02	≈100	500/2000

After initialisation of the graph with appropriate parameters, the nodes are restricted to be source nodes and destination nodes. The nodes representing user equipments (UEs) are grouped as source nodes (S) and the nodes representing the data center nodes are grouped as destination nodes (Σ).

The three different classes of services are defined by their requirements of demanded capacity, required reliability, tolerable latency, priority, etc. Any i^{th} service that arrives to the system is identified by a quadruple (s_i,σ_i,α_i,D_i) , where s_i is the source node for the i^{th} service, σ_i is the destination node, ,α_i is the indicator to indicate if the service is elastic (α_i =1) or inelastic (α_i=0). D_i is the demand vector which specifies the capacity, latency, reliability and priority demands of the arriving service and is denoted by {[c_{min}, c_{max}], [SL_{min}, SL_{max}], [SR_{min}, SR_{max}], SP]}, if the service is elastic or by {[c, SL,SR, SP]}, if the service is inelastic. Since the elastic type of service can endure a range of values for any parameter , the demand set is specified in minimum and maximum terms.

The time scale is taken as 24 hours to represent a real day situation and the three type of services arrive in a Poisson distributed manner to the system which are then sent to the transmission queue. The traffic arrival distribution over a period of 24 hours is shown in Figure 4.4.

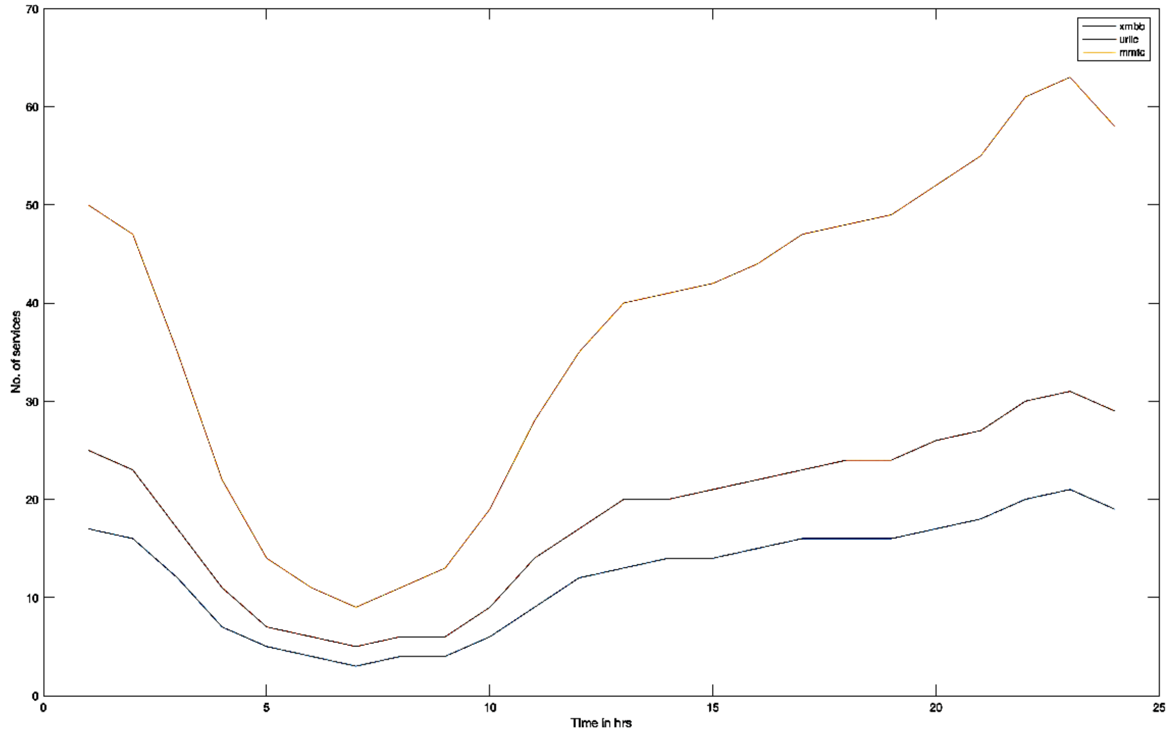


Figure 4.4 Distribution of Traffic of Three Types of Services Over 24 Hours

Figure 4.5 shows the block diagram of the incoming services and scheduling. Figure 4.5 shows the flowchart of the algorithm.

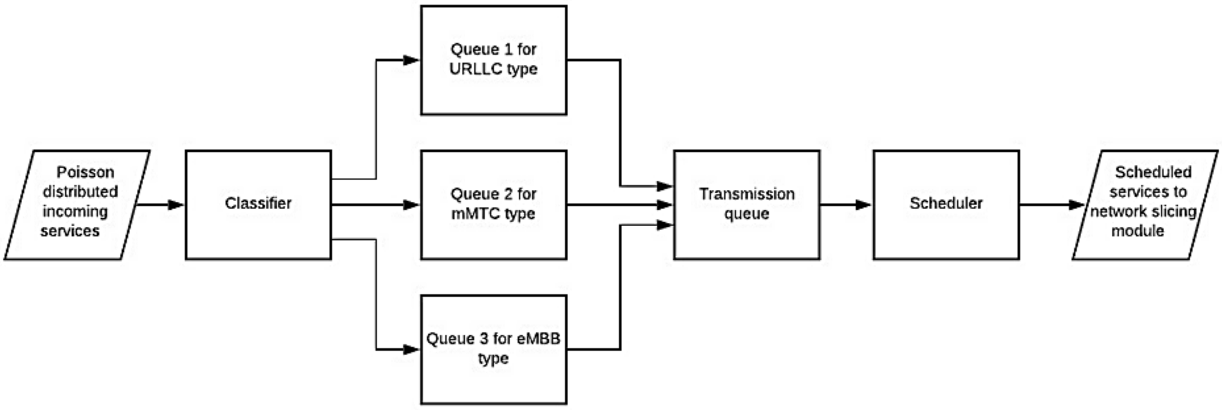


Figure 4.5 Block Diagram of Service Arrival and Scheduling Process

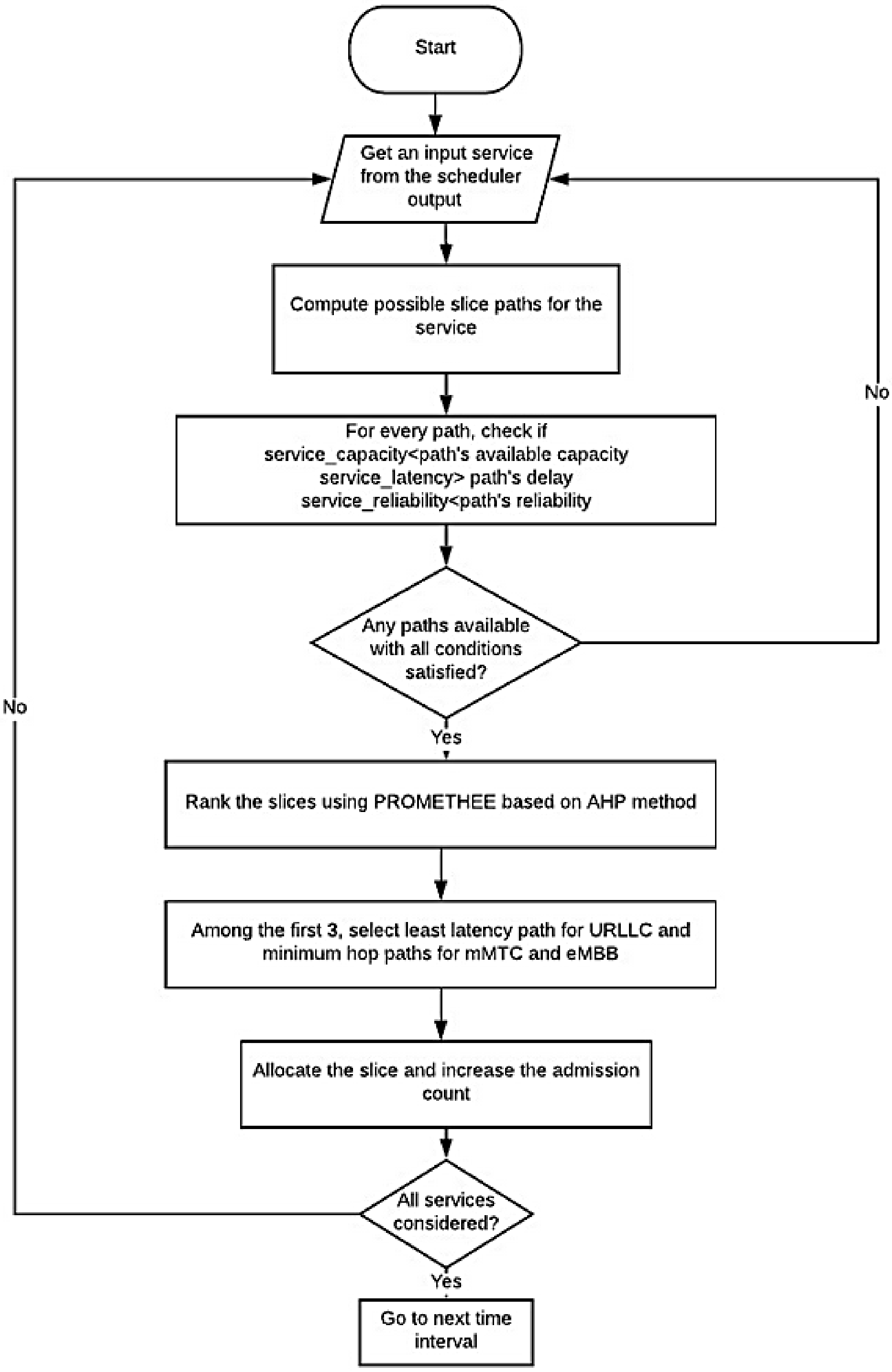


Figure 4.6 Flowchart of Network Slicing Using PROMETHEE Method based on AHP

The Poisson distributed services arrive and the classifier classifies them according to the type of service i.e. eMBB, mMTC or URLLC. From these three queues, the services are put into the transmission queue in a Round-Robin fashion. Like how any scheduling process cannot be instantaneous that just sends every other service that arrives, here also, the services wait in the transmission queue until all the services for that interval are put in the queue. But since the services wait in the transmission queue for a little while and since the data on the lifetime of a service is not available, a parameter representing a equivalence of waiting time is coded and used in MATLAB.

The scheduler now decides the order in which to send the arrived services into the network slicing module. Some techniques based on FCFS, round-robin, priority,etc. were dicussed in Chapter 2 of the literature review and the disadvantages were mentioned in Chapter 3. To overcome the disadvantages of too much prioritisation by some services and starvation by others it is necessary that all the criteria are weighed upon before allowing the service for subsequent slice allocation. For example, URLLC might suffer in FCFS scheme but preferentially gains acceptance in priority based scheme due to its low delay-tolerance nature, whereas eMBB might have more chance of getting accepted in FCFS (whenever it happens to arrive) than the priority one. Therefore, any particular type of service will always be at disadvantage and the solution is always a compromise, as observed in the literature review chapter. Therefore, this becomes a valid reason to adopt MADM methods in the scheduling as well as slicing scenarios.

The MADM related works discussed in the literature review considered the application of the techniques only in the slicing domain. In this work, the application is also considered in the beginning scheduling stage itself because the manner of scheduling has an impact on how the network resources will be used since whatever service goes inside the system consumes resources and then the resource de-allocation would depend on when the service's lifetime/task is is over.

Using the improved PROMETHEE combined with AHP, the services' scheduling order is decided while taking into account the relative importance of each of the service parameters/criteria involved. The parameters used at this step are values from the demand vectors of the incoming services that are the Service Reliability (SR), Service Latency (SL), Service Priority (SP), and Time in System (TS). TS is measured as the difference between the time the service enters the transmission queue and the point of time when it is considered for scheduling along with others. An example of how the incoming services are aligned so as to be considered for network slicing using the improved PROMETHEE combined with AHP method is explained in APPENDIX 1.

Once the scheduling module determines the service order each service is sent sequentially to the network slicing module. Figure 4.6 depicts the starting of this step. In this module, the network parameters are worked upon. As a general idea, the scheduling procedure works with the service parameters and the slicing module focusses on the network parameters. All possible paths between the source and destination node is computed for every incoming service and for each path, the conditions given in Eqs. (4.4) to (4.6) are checked.

$$capacity_{service} \leq \sum_{i=1}^{all\ links\ of\ the\ path}(Available\ capacity\ on\ link_i) \tag{4.4}$$

$$latency_{service} \leq \sum_{i=1}^{all\ links\ of\ the\ path}(delay\ of\ link_i) \tag{4.5}$$

$$reliability_{service} \leq \sum_{i=1}^{all\ links\ of\ the\ path}(Reliability\ of\ link_i) \tag{4.6}$$

If all the three conditions are satisfied then the path is considered eligible as an input to the PROMETHEE method subsequently. Out of all possible paths, only some paths would satisfy all the three requirements and be considered for ranking by the PROMETHEE method. Since the three service types are specifically sensitive to one or more *network* paramters, the weights to the minimum available capacity on path, delay and reliability of the *network* for the implementation of PROMETHEE method are taken as mentioned. For eMBB, since it needs high capacity, the weightages assigned to the minimum available path capacity (Ca), path delay (D) and the path reliability (R) are 0.60, 0.20 and 0.20, respectively. For URLLC, which is latency sensitive but also requiring a reliable slice, the weights to Ca, D and R are assigned as 0.20,0.20, and 0.40, respectively. For mMTC the weights of Ca, D, R are taken as 0.40, 0.30 and 0.30, respectively.

After the PROMETHEE method gives ranks to the slice alternatives, instead of selecting the first path, the minimum delay path among the first few highly ranked alternatives is allocated to the slice request of URLLC type, the minimum hop path for the mMTC and eMBB type traafic. This extra step is to ensure that a best path/slice suited for each service type is chosen. It would be intuitive to say to choose the minimum delay or hop paths directly without the MADM application but it is to be noted that the inclusion of MADM makes sure that the other two parameters are also given due importance when deciding a possible slice for the request since it is not always true that the least delay or hop path is the best in terms of fulfilling, say, capacity requirements too.

The next chapter presents the results that are generated based on the methodology presented in this chapter and the related discussion.

Chapter 5

RESULTS AND DISCUSSION

This chapter presents the results and the discussion on them. The output parameters mainly considered to measure the performance of the proposed scheduling and the slicing process are service admission probability, admission ratio and mean link utilization. The admission probability, admission ratio and the mean link utilisation at time t are given by Eqns. (5.1) to (5.3).

$$\text{Prob\_ad (t)} = \frac{total\ admitted\ services}{total\ services\ that\ arrived} \tag{5.1}$$

$$admission\ ratio\ (t) = \frac{number\ of\ admitted\ services}{number\ of\ rejected\ services} \tag{5.2}$$

$$Mean\ link\ utilisation(t) = mean\frac{(Original\ capacity - utilised\ capacity)}{original\ capacity} \tag{5.3}$$

Figure 5.1 shows the admission probability of the services arriving at the network over a 24 hours period in a Poisson distributed manner. It is observed from the graph that even though the probability has fluctuated slightly, but overall it has been efficient and fair in accepting the services.

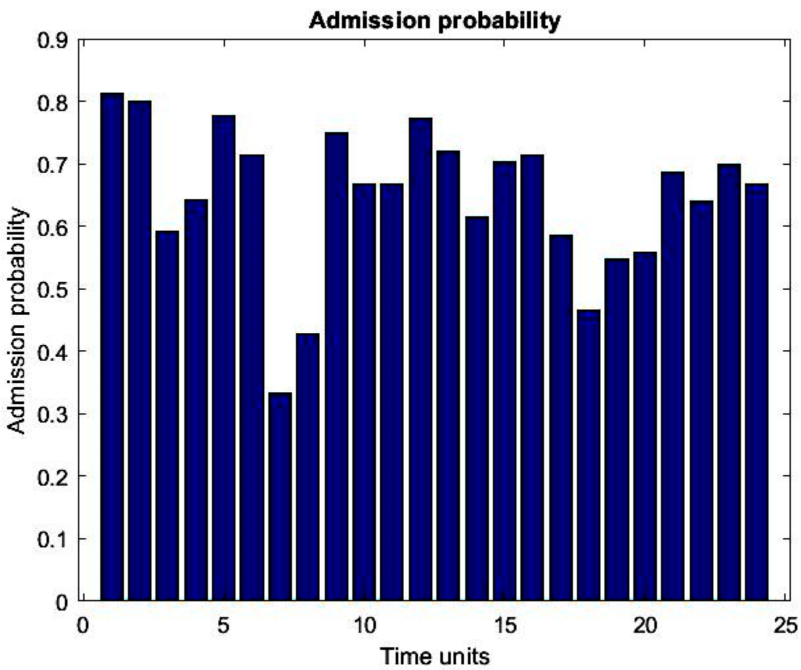


Figure 5.1 Admission Probability of the All Services Together Over 24 Hours

This observation can be supported by observing Figure 5.2 and Figure 5.3. Figure 5.2 shows the total accepted services of all types over all 24 hours. It is evident from the graph that the admissions have gradually increased with corresponding minimum rejections. There seems to be a decline in service acceptance but it is due to the reduced arrivals in that interval rather than rejections. This is evident in Figure 5.3 since there is no increase in rejection to account for the decline in acceptance.

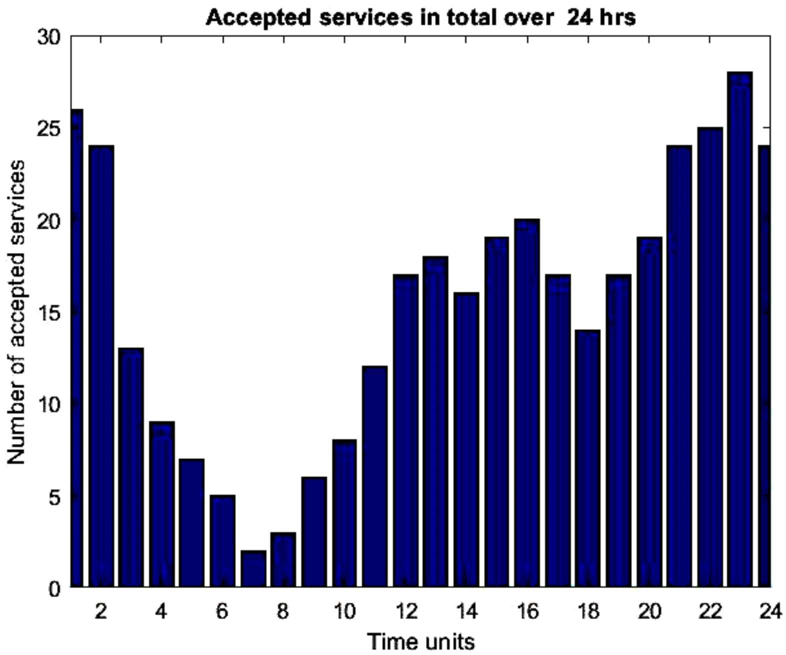


Figure 5.2 Accepted Services Over 24 Hours

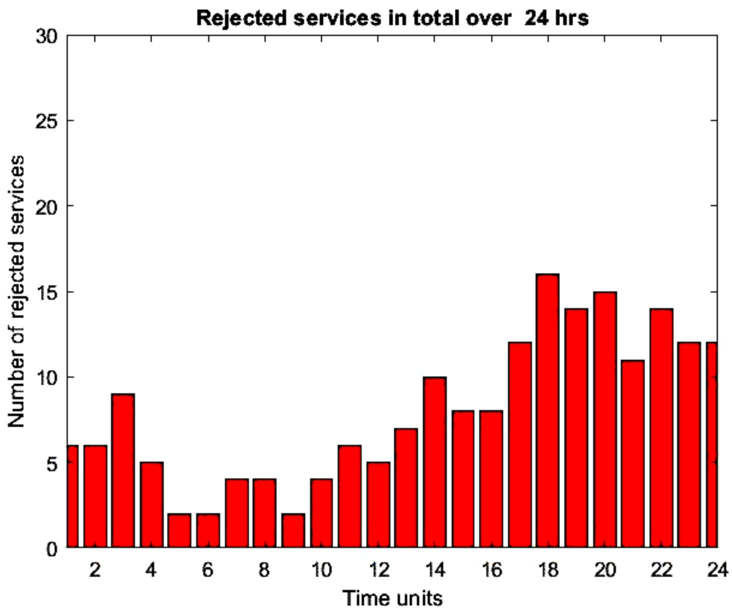


Figure 5.3 Rejected Services Over 24 Hours

From Figure 5.3, it can be seen that the rejections also increased over time but are less than the accepted ones. The increase in rejections is due to the fact that the capacity resource decreases over time and the deallocation depends on the departures of the services.

Figure 5.4 depicts the admission ratio which is calculated using Eq. (5.2) at each time t. For the analysis of this graph, an average admission ratio is found out by taking the average of all the ratios over the 24 hours period and it is noted that, on average, the accepted services are twice the rejected ones.

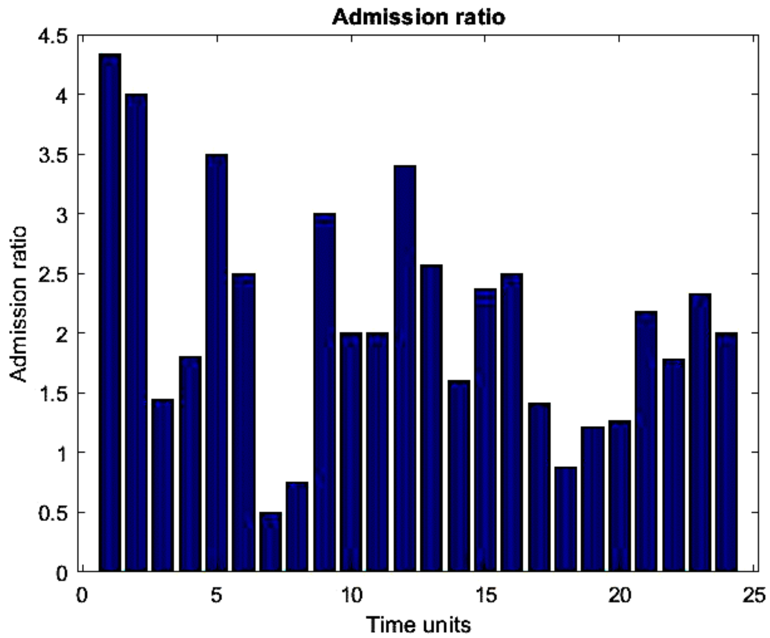


Figure 5.4 Admission ratio

Figure 5.5 shows the mean link utilization of the network, i.e. average of the usage of all links in the network on a whole. It is observed that the utilization has improved over time. The average utilisation over the 24 hours period is 81.27%. The increase in utilization is a good indication since the link utilization is directly related to the number of services active in the system. From Figure 5.2, it can be seen that the accepted number of services increased over time and consequently the usage of links also increased leading to increased mean utilization as shown in Figure 5.5.

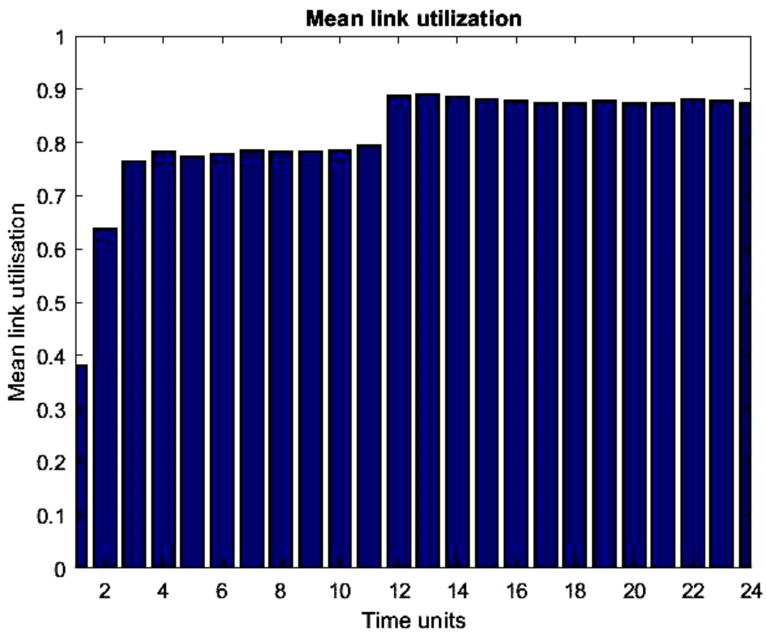
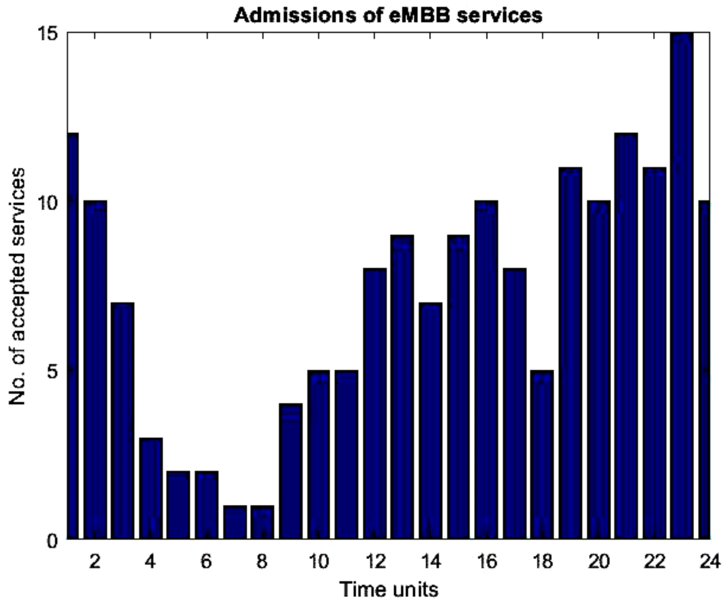


Figure 5.5 Mean Link Utilization

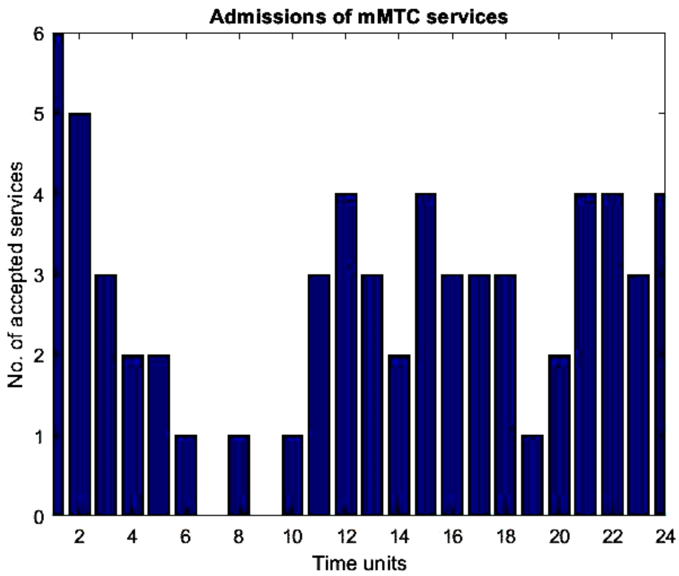
From Figure 5.5, it can be observed that the links are well utilized and therefore the new services find it difficult to satisfy their demands and get rejected. To solve this issue, either the departures have to be made or pre-emptive interruption should be done to stop some of the low priority services and allocate resources to the incoming high priority ones or incoming services should be reduced in the beginning itself to reduce the load. But departures cannot be initiated just for the cause of reducing rejections and interruption and reduction in number of input services also means rejection of the services though at the beginning.

From Figure 5.5, it also seems that there are no points where there is zero rejection. But seeing the rejection graphs of each type of service in Figure 5.7, it can be noted that there have been times when there were no rejections in one or two service types but a rejection of third type of service caused rejection that led to presence of a rejection in the total rejections graph of Figure 5.5. It is to be noted that even though capacity is the only changing resource, a service can get rejected even if it fails to satisfy the delay and reliability condition. That is why, working along with the network congestion issue from the perspective of the resource usage might still lead to rejections.

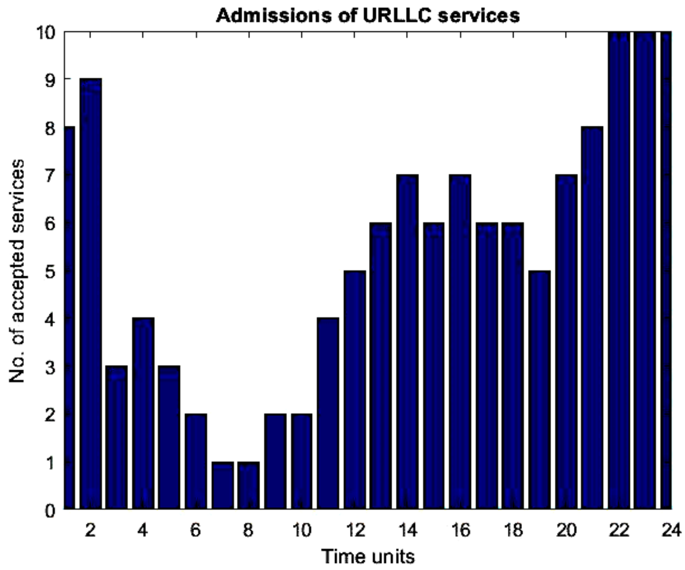
Figures 5.6 and 5.7 show the accepted and rejected services of eMBB, mMTC and URLLC. The common observation is that the number of rejections is less than the admissions even if rejections occur in parallel with the admissions. This observation also confirms the behavior of all the graphs mentioned before in terms of fair admission probability, the conclusion from the admission ratio graph and the mean link utilization graph. The decrease in the admissions for a small period in the beginning is due to less number of arrivals. The pattern of the graphs follows the pattern of the input traffic graph of Figure 4.4.



(a)

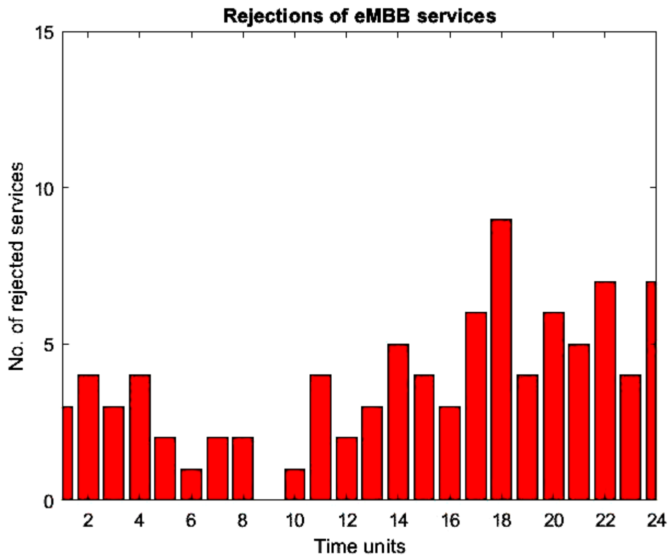


(b)

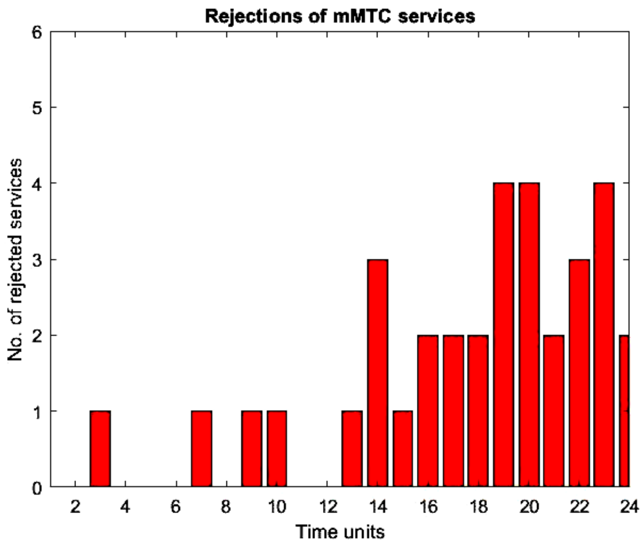


(c)

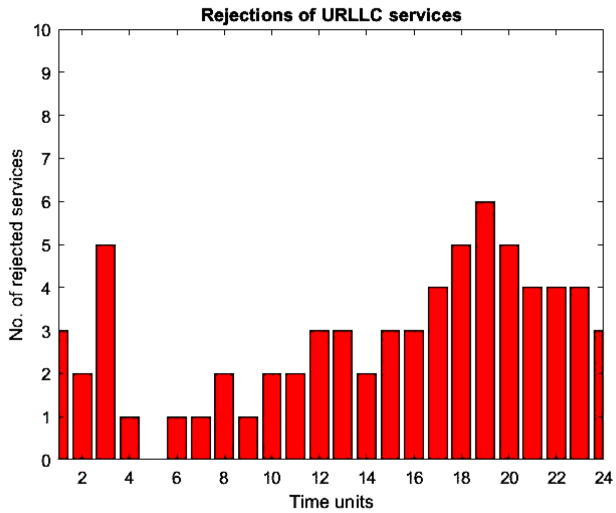
Figure 5.6 Admitted Services of (a) eMBB, (b) mMTC, and (c) URLLC Type of Service



(a)



(b)



(c)

Figure 5.7 Rejected Services of (a) eMBB, (b) mMTC, and (c) URLLC Type of Service

As mentioned in the chapter on literature review, very few works are available that make use of MADM techniques to *5G* problems. The previous researchers had used TOPSIS and VIKOR techniques in a different scenario like the number and distribution of services, values of node and link parameters, number of network resources, etc. Since the implementation scenarios are different, it is not fair to compare our results with the results of these works directly. Hence, in this work, the results obtained by the proposed MADM technique are analyzed independently. It is to be noted that the TOPSIS and VIKOR methods have the drawbacks of being computationally intensive, difficult to understand and not having a systematic way of deciding the weights of the criteria. Assigning wrong weights or inconsistent weights changes the solution and will not be optimal anymore. However, the proposed improved PROMETHEE based on AHP overcomes these drawbacks. The proposed method is computationally inexpensive and the weights of the criteria computed by AHP are proved to be valid and consistent for the problem considered. The consistency ratio of the judgments of relative importance of the criteria should be less than or equal to 0.1 and the weights used in this work produce a consistency ratio value of 0.027 (shown in Appendix 1).

The results show that it is possible to make use of PROMETHEE method based on AHP as a multi-attribute decision making technique for scheduling and network slicing solutions as an alternative to multi-objective decision making techniques (MODM). As mentioned in the preceding chapters, this approach takes into account various criteria of the services and the network to determine the service to allow (in an environment of multiple services with conflicting demands) for the slicing as well as to find the best possible slice for a given service.

The next chapter presents the conclusions of the thesis.

Chapter 6
CONCLUSIONS

Network slicing technology helps 5G networks to achieve the aim of serving diverse applications. It enables operators to partition a single physical network into various slices for different traffic types where each slice has an isolated set of network resources. Slicing is closely related to scheduling. When the number of requests is less and the requests have homogeneous network requirements, slicing is easy and straightforward. However, with a large number of requests with diverse demands (like eMBB, mMTC and URLLC), slicing needs to be performed by considering both the demands and the network resource constraints. This work aims to contribute a solution to the end-to-end slicing problem in C-RAN by the use of the MADM techniques, namely improved PROMETHEE method combined with the AHP.

A full network topology consisting of RAN, microwave backhaul, fibre backhaul and data center network is simulated and different requests arrive in a Poisson based manner over 24 hours. The requests are scheduled using the improved PROMETHEE method combined with AHP by considering their tolerable latency, required reliability, priority and waiting time in the queue. The weights of these criteria are found using the AHP method and the ranking of alternative services is done using the PROMETHEE method. The scheduled services are considered for slicing, and the slicing is performed by comparing the service requirements with the then-existing network resources to see if a path/slice is feasible for the arrived request. The improved PROMETHEE combined with the AHP method is applied to rank the slices and a minimum latency path or a minimum hop path among the first few ranked slices is selected. The ranking is performed to choose the best possible slice for the particular type of request based on what network resource is vital for it.

The results show that the proposed method can effectively handle different 5G traffic requests under varying load conditions over time. The admission probability obtained increases over time with little fluctuations in the beginning due to a decrease in arrivals itself rather than rejections. The admission ratio results show that, on an average, the average number of accepted services is twice the rejected ones. Consequently, the mean link utilization also increases over time. Thus, the results prove that the improved PROMETHEE combined with the AHP method for scheduling and network slicing problems is useful and feasible. The future works can couple this work with the issue of optimal placement of virtual network functions (VNF) too.

APPENDIX 1

To demonstrate the step-by-step working of the improved PROMETHEE method, an example is presented here.

Step 1: Let the attributes identified are Service Reliability (SR), Service Latency (SL), Service Priority (SP), and Time in System (TS). Let $eMBB_1$, $eMBB_2$, $mMTC_1$, URLLC, and $mMTC_2$ are different services that arrived with their requirements of reliability, latency, priority and time in system parameter that is measured as the difference between the time when scheduling starts and the time of arrival time into the transmisison queue. In fact, the same type of service can arrive or appear again. Let eMBB has arrived twice with different requirements, then it is named as $eMBB_1$ and $eMBB_2$. eMBB is an elastic type of service and its demands vary over time in a fixed range. Similarly, let mMTC has arrived twice with different requirements, then it is named as $mMTC_1$ and $mMTC_2$. In actual practice, the number of input services to the network slicing module will be huge in muber. This example is just for demonstration only. More than 100 services can be there at a time in actual practice. Table A1.1 shows the decision table containing the information about the alternatives, attributes and the corresponding data. Higher values of SR and SP and lower values of SL and TS are desired.

Table A1.1 Decision Table

Service Alternatives	**Attributes**			
	Service Reliability (SR)	Service Latency (SL)	Service Priority (SP)	Time in System (TS)
$eMBB_1$	0.90	100	0.15	4
$eMBB_2$	0.85	110	0.15	3
$mMTC_1$	0.90	50	0.25	2
URLLC	0.99	10	0.60	1
$mMTC_2$	0.90	50	0.25	0

Step 2: Decide the attributes' weights

The four attributes considered asre SR, SL, SP and TS. The procedure for finding the weights is outlined below.

Step 2.1: Find the relative importance of attributes

Let the attribute SR is considered slightly more important than SL. Then the relative importance of 3 is assigned corresponding to SR vs. SL. The reciprocal of 3, i.e. 1/3 is assigned to SL vs. SR. The attribute SP is considered strongly more important than SR. Hence relative importance of 5 is assigned corresponding to SP vs. SR. The reciprocal of 5, i.e. 1/5 is assigned to SR vs. SP. Similarly, the other relative importance relations are prepared. It may be noted that these relative importance relations assigned are just for demonstration only. In actual practice, the designer or decision maker can decide the relative importance relations based on his requirements. The relative importance marix $M1_{4x4}$ is prepared as,

$$
\begin{array}{cc}
\text{Attribute} & \begin{array}{cccc} \text{SR} & \text{SL} & \text{SP} & \text{TS} \end{array} \\
\begin{array}{c} \text{SR} \\ \text{SL} \\ \text{SP} \\ \text{TS} \end{array} &
\begin{bmatrix} 1 & 3 & 1/5 & 1/5 \\ 1/3 & 1 & 1/7 & 1/7 \\ 5 & 7 & 1 & 1 \\ 5 & 7 & 1 & 1 \end{bmatrix}
\end{array}
$$

Step 2.1: Relative normalized weights (w_j)

The geometric means are calculated using Eq. (3.2) and the weights are calculated using Eq. (3.3) of section 3 and these are,

$GM_{SR} = (1*3*1/5*1/5)^{1/4} = 0.588$

$GM_{SL} = (1/3*1*1/7*1/7)^{1/4} = 0.287$

$GM_{SP} = (5*7*1*1)^{1/4} = 2.432$

$GM_{TS} = (1*3*1/5*1/5)^{1/4} = 2.432$

$\sum GM_j = 5.739$

And

$w_{SR} = 0.588/5.739 = 0.1025$

$w_{SL} = 0.287/5.739 = 0.0500$

$w_{SP} = 2.432/5.739 = 0.4237$

$w_{TS} = 2.432/5.739 = 0.4237$

Step 2.2: The matrix M3 is computed as M1*M2.

$$\begin{bmatrix} 1 & 3 & 1/5 & 1/5 \\ 1/3 & 1 & 1/7 & 1/7 \\ 5 & 7 & 1 & 1 \\ 5 & 7 & 1 & 1 \end{bmatrix} * \begin{bmatrix} 0.1025 \\ 0.0500 \\ 0.4237 \\ 0.4237 \end{bmatrix} = \begin{bmatrix} 0.4219 \\ 0.2051 \\ 1.7099 \\ 1.7099 \end{bmatrix} = \mathrm{M3}$$

The matrix M4 is computed as M3/M2.

$$\begin{bmatrix} 0.4219/0.1025 \\ 0.2051/0.0500 \\ 1.7099/0.4237 \\ 1.7099/0.4237 \end{bmatrix} \begin{matrix} = \\ = \\ = \\ = \end{matrix} \begin{bmatrix} 4.116 \\ 4.102 \\ 4.0356 \\ 4.0356 \end{bmatrix} = \mathrm{M4}$$

Step 2.3: The maximum Eigen value (λ_{max}) is the average of matrix M4, i.e. 4.072

Step 2.4: The consistency index CI = (4.072 - 4)/(4 - 1) = 0.0241.

Step 2.5: The consistency ratio CR = 0.0241/0.89 = 0.027. The calculated value of CR is 0.027 and it is much less than the allowed value of 0.1. Hence, it can be said that there is good consistency in the relative importance relations assigned in M1.

Step 3: Preference function calculation (P_j)

The preference values P_j resulting from the pairwise comparisons of the alternatives with respect to attributes are given below. When an alternative *a*1 dominates another alternative *a*2 then 1 is assigned corresponding to *a*1 vs. *a*2, and 0 is assigned corresponding to *a*2 vs. *a*1. However, if two alternatives are equal then no dominance exists and 0 is assigned in both the cases.

SR (0.1025)	$eMBB_1$	$eMBB_2$	$mMTC_1$	URLLC	$mMTC_2$
$eMBB_1$	---	1	0	0	0
$eMBB_2$	0	---	0	0	0
$mMTC_1$	0	1	---	0	0
URLLC	1	1	1	---	1
$mMTC_2$	0	1	0	0	---

SL (0.0500)	$eMBB_1$	$eMBB_2$	$mMTC_1$	URLLC	$mMTC_2$
$eMBB_1$	---	1	0	0	0
$eMBB_2$	0	---	0	0	0
$mMTC_1$	1	1	---	0	0
URLLC	1	1	1	---	1
$mMTC_2$	1	1	0	0	---

SP (0.4237)	$eMBB_1$	$eMBB_2$	$mMTC_1$	URLLC	$mMTC_2$
$eMBB_1$	---	0	0	0	0
$eMBB_2$	0	---	0	0	0
$mMTC_1$	1	1	---	0	0
URLLC	1	1	1	---	1
$mMTC_2$	1	1	0	0	---

TS (0.4237)	$eMBB_1$	$eMBB_2$	$mMTC_1$	URLLC	$mMTC_2$
$eMBB_1$	---	0	0	0	0
$eMBB_2$	1	---	0	0	0
$mMTC_1$	1	1	---	0	0
URLLC	1	1	1	---	0
$mMTC_2$	1	1	1	1	---

The resulting preference indices are shown below.

	$eMBB_1$	$eMBB_2$	$mMTC_1$	URLLC	$mMTC_2$
$eMBB_1$	---	0.1525	0	0	0
$eMBB_2$	0.4237	---	0	0	0
$mMTC_1$	0.8974	1	---	0	0
URLLC	1	1	1	---	0.5762
$mMTC_2$	0.8974	1	0.4237	0.4237	---

Step 4: Calculation of leaving ,entering flow and net flow

$\Phi^{+}(a)$ = 0.1525, 0.4237, 1.8974, 3.5762, and 2.7445

$\Phi^{-}(a)$ = 3.2185, 3.1525, 1.4237, 0.4237 and 0.5762

$\Phi(a) = \Phi^{+}(a) - \Phi^{-}(a)$ = -3.066, -2.7288, 0.4737, 3.1525 and 2.1683

From the values of net flow, the order of services that should proceed for network slicing are in the following order:

URLCC, $mMTC_2$, $mMTC_1$, $eMBB_2$ and $eMBB_1$.

REFERENCES

[1] E. Ezhilarasan and M. Dinakaran, "A Review on Mobile Technologies: 3G, 4G and 5G," In *Proceedings – 2nd International Conference on Recent Trends and Challenges in Computational Models,* Tindivanam, 3-4 Feb., 2017, pp. 369-373.

[2] A. Shaik, R. Borgaonkar, S. Park, and J. P. Seifert, "New Vulnerabilities in 4G and 5G Cellular Access Network Protocols: Exposing Device Capabilities," In *Proceedings - 12th Conference on Security and Privacy in Wireless and Mobile Networks,* Miami, May 2009, pp. 221-231.

[3] O. Idowu-Bismark, K. O. Okokpujie, H. Ryan, and M. O. Adedokun, "5G Wireless Communication Network Architecture and its Key Enabling Technologies," *Int. Rev. Aerospace Eng.*, vol. 12, no. 2, pp. 70-82, 2019.

[4] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network Slicing in 5G: Survey and Challenges. *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94-100, 2017.

[5] T. Guo, and A. Suárez, "Enabling 5G RAN Slicing with EDF Slice Scheduling," *IEEE Trans. Vehicular Technol.*, vol. 68, no. 3, pp. 2865-2877, 2019.

[6] Q. Jia, R. Xie, T. Huang, J. Liu, and Y. Liu, "Caching Resource Sharing for Network Slicing in 5G Core Network: A Game Theoretic Approach," *J. Organiz. End User Comput.*, vol. 31, no. 4, pp. 1-18, 2019.

[7] L. Tang, Q. Tan, Y. Shi, C. Wang, and q. Chen, "Adaptive Virtual Resource Allocation in 5G Network Slicing Using Constrained Markov Decision Process," *IEEE Access*, vol. 6, no. 6, pp. 61184-61195, 2018.

[8] B. Han, V. Sciancalepore, D. Feng, X. Costa-Perez, and H. D. Schotten, "A Utility-Driven Multi-Queue Admission Control Solution for Network Slicing," in *Proceedings - IEEE Conference on Computer Communications,* Paris, 29 April - 2 May, 2019, pp. 55–63.

[9] M. R. Raza, C. Natalino, P. Öhlen, L. Wosinska, and P. Monti, "A Slice Admission Policy Based on Reinforcement Learning for a 5G Flexible RAN," in *Proceedings - European Conference on Optical Communication,* Rome, 23-27 September, 2018.

[10] R. Challa, V. Zalyubovskiy, S. M. Raza, H. Choo, and A. De, "Network slice admission

model: Tradeoff between monetization and rejections," *IEEE Systems J.*, vol. 14, no. 1, pp. 657–660, 2020.

[11] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, and X. Costa-Perez, "A Machine Learning Approach to 5G Infrastructure Market Optimization," *IEEE Trans. Mob. Comput.*, vol. 19, no. 3, pp. 498–512, 2020.

[12] A. Anand, G. De Veciana, and S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, 2020.

[13] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, B. Ottersten, "Slicing based resource allocation for multiplexing of eMBB and URLLC services in 5G wireless networks," in *Proceedings - IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks*, Limassol, 11-13 September, 2019, pp. 1–5.

[14] N. Ginige, K. Manosha, N. Rajatheva, Latva-aho Matti - 2020 IEEE 91st, and U. 2020, "Admission Control in 5G Networks for the Coexistence of eMBB-URLLC Users," in *Proceedings - IEEE 91st Vehicular Technology Conference*, 2020, pp. 1–6.

[15] H. D. R. Albonda and J. Perez-Romero, "An Efficient RAN Slicing Strategy for a Heterogeneous Network with eMBB and V2X Services," *IEEE Access*, vol. 7, pp. 44771–44782, 2019.

[16] D. Dietrich, C. Papagianni, P. Papadimitriou, and J. S. Baras, "Network function placement on virtualized cellular cores," in *Proceedings - 9th International Conference on Communication Systems and Networks,* Bangalore, 4-8 June, 2017, pp. 259–266.

[17] D. Sattar and A. Matrawy, "Optimal Slice Allocation in 5G Core Networks," *IEEE Netw. Lett.*, vol. 1, no. 2, 48–51, 2019.

[18] A. Nakao, P. Du, Y. Kiriha, F. Granelli, A. A. Gebremariam, T. Taleb, and M. Bagga, "End-to-end network slicing for 5G mobile networks," *J. Inf. Process.*, vol. 25, pp. 153–163, 2017.

[19] S. Vassilaras, Z. Gkatikis, N. Liakopoulos, I. N. Stiakogiannakis, M. Qi, L. Shi, L. Liu, M. Debbag, and G. S. Peschos, "The Algorithmic Aspects of Network Slicing," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 112–119, 2017.

[20] A. Kammoun, N. Tabbane, G. Diaz, A. Dandoush, and N. Achir, "End-to-end efficient heuristic algorithm for 5G network slicing," in *Proceedings - International Conference on Advanced Information Networking and Applications,* Krakow, 16-18 May, 2018, pp. 386–392.

[21] R. Bassoli and F. Granelli, "An algebraic approach to network slicing," in *Proceedings- 25th European Wireless 2019 Conference,* Aarhus, 2019, pp. 55–60.
[22] B. Bakmaj, Z. Bojkovic, and M. Bakmaj, "TOPSIS-based reference for network slice selection in 5G mobile systems," *International Journal of Communication Systems*, vol. 33, no. 11, e4395, 2020.
[23] R. V. Rosa and C. E. Rothenberg, "The Pandora of Network Slicing: A Multi-Ctriteria Analysis," *Trans. Emerg. Telecomm. Technol.,* vol. 31, no.1, e3651, 2020.
[24] X. Li, C. Guo, L. Gupta, and R. Jain, "Efficient and Secure 5G Core Network Slice Provisioning Based on VIKOR Approach," *IEEE Access*, vol. 4, pp. 1–13, 2019.
[25] Y. Yu, B. Yong, and C. Lan, "Utility-dependent network selection using MADM in heterogeneous wireless networks," in *Proceedings-IEEE International Symposium on Personal, Indoor and Mobile Radio Communications,* Athens, 3-7 September, 2007.
[26] J. D. Martínez-Morales, U. Pineda-Rico, and E. Stevens-Navarro, "Performance comparison between MADM algorithms for vertical handoff in 4G networks," in *Proceedings- 7th International Conference on Electrical Engineering, Computing Science and Automatic Control,* Tuxtla Gutierrez, 8-10 September, 2010, pp. 309–314.
[27] D. Maroua, O. Mohammed, and A. Driss, "VIKOR for multi-criteria network selection in heterogeneous wireless networks," in *Proceedings - International Conference on Wireless Networks and Mobile Communications,* Fez, 26-29 October, 2016, pp. 82–86.
[28] K. Savitha and C. Chandrasekar, "Trusted Network Selection using SAW and TOPSIS Algorithms for Heterogeneous Wireless Networks," *Int. J. Comput. Appl.*, vol. 26, no. 8, pp. 22–29, 2011.
[29] L. Mohamed, C. Leghris, and A. Abdellah, "A survey and comparison study on weighting algorithms for access network selection," in *Proceedings - 9th Annual Conference on Wireless On-Demand Network Systems and Services,* Courmayeur, 9-11 January, 2012, pp. 35–38.
[30] R. V. Rao, *Decision Making in the Manufacturing Environment: Using Graph Theory and Fuzzy Multiple Attribute Decision Making Methods*, Springer, 2007.
[31] R. V. Rao, *Decision Making in the Manufacturing Environment Using Graph Theory and Fuzzy Multiple Attribute Decision Making Methods - Volume 2*, Springer, 2013.
[32] P. C. Fishburn, "Additive Utilities with Incomplete Product Set: Applications to Priorities and sharings," *Oper. Res. Soc. Am. (ORSA), Balt.* 1967.
[33] T. L. Saaty, "*The Analytic Hierarchy Process*," McGraw Hill, New York, 1988.
[34] G. H. Tzeng and J. J. Huang *"Multiple Attribute Decision Making: Methods and*

Applications", CRC Press, Boca Raton, 2011.

[35] S. Opricovic and G. H. Tzeng, "Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS," *Eur. J. Oper. Res.*, vol. 156, no. 2, pp. 445–455, 2004.

[36] J. P. Brans, P. Vincke, and B. Mareschal, "How to select and how to rank projects: The PROMETHEE method," *Eur. J. Oper. Res.*, vol. 24, no. 2, pp. 228–238, 1986.

[37] R. V. Rao and B. K. Patel, "Decision making in the manufacturing environment using an improved PROMETHEE method," *Int. J. Prod. Res.*, vol. 48, no. 16, pp. 4665–4682, 2010.